An Illumination Effect Descriptor for Video Sequences

Jürgen Stauder

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

An Illumination Effect Descriptor for Video Sequences

Jürgen Stauder

No 3866

January 2000

THÈME 3

ISSN 0249-6399 ISRN INRIA/RR--3866--FR+ENG

An Illumination Effect Descriptor for Video Sequences

Jürgen Stauder

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Temics

Rapport de recherche n° 3866 — January 2000 — 20 pages

Abstract: In the context of indexing of video data bases, the future standard MPEG-7 will provide descriptors for motion, shape, texture and color to characterize and identify video scenes. This report presents a descriptor, which has been proposed to MPEG-7 recently, for effects caused by the scene illumination. The proposed descriptor addresses temporal changes of object shading, of cast shadows and of the global illumination intensity by one single scalar. It is based on temporal changes of image luminance along motion trajectories. Further, an automatic, noise adaptive method for the extraction of the illumination effect descriptor from a video sequence is presented. A performance analysis and sample retrieval experiments show that extracted descriptor values are sufficiently precise and sensible to distinguish video sequences with different scene illumination.

Key-words: Video databases, image retrieval, indexing, illumination, shading

(Résumé : tsvp)

Supported by a European Marie Curie Research Training Grant

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : 02 99 84 71 00 - International : +33 2 99 84 71 00
Télécopie : 02 99 84 71 71 - International : +33 2 99 84 71 71

Un descripteur d'effets d'illumination pour les séquences vidéo

Résumé : La mise en oeuvre de méthodes d'indexation des bases de données vidéo est importante pour en permettre une gestion efficace. Dans ce cadre, le futur standard MPEG-7 prévoira des descripteurs pour le mouvement, la texture et la couleur afin d'identifier et de caractériser des scènes vidéo. Dans ce rapport, un descripteur pour les effets d'illumination est présenté, qui était récemment proposé dans le cadre de la normalisation MPEG7. Ce descripteur permet de caractériser, grâce à un seul paramètre, certaines propriétés liées au changement temporel de l'ombrage, des ombres portées et de l'intensité de l'illumination globale de la scène. Ce paramètre est basé sur le changement temporel de la luminance le long des trajectoires de mouvement. De plus, une méthode automatique et adaptative au bruit pour l'extraction de ce descripteur à partir d'une séquence vidéo est présentée. Une analyse de l'efficacité et des expériences de recherche dans une base de données vidéo montrent que les valeurs de descripteur extraites sont assez précises et sensibles pour distinguer des scènes vidéo avec de illuminations différentes.

Mots-clé : Base de données vidéo, recherche, indexation, illumination, ombrage

1 Introduction

In the emerging future market of storage and retrieval of video sequences in huge data bases the problem of video indexing arises. In this context, the future standard MPEG-7 [11] is being prepared. For video indexing, MPEG-7 will define descriptors for the content of the video sequences based on motion, texture, color and shape. The video sequences together with the descriptors form an indexed data base. The access to an indexed data base can then be supported by a search machine that selects only those sequences containing scenes with demanded motion, texture, color and shape.

If the illumination of the video scene is not diffuse, illumination effects may occur in the video sequences. Illumination effects are cast shadows [25][28], shading [22], interreflections [17], specular reflections [8] and transparency [14][6]. These effects may disturb the indexing process where motion, texture, color and shape information should be extracted from a video sequence. In the computer vision literature, one idea is to avoid illumination effects by an optimal arrangement of light sources in the video scene [37][15]. But in video indexing, the video sequences are already present and can not be influenced. Another idea is to extract illumination invariants from an image [16][35] to avoid the influence of scene illumination.

This report [33][34] does not regard illumination effects as disturbing. We go one step further and regard illumination effects as useful information about the scene illumination. Unfortunately, neither the descriptors presented in the literature [23][2] nor those currently discussed for MPEG-7 [26] do regard the *illumination* of the video scene. For the first time, this report describes an illumination effect descriptor. The descriptor has been proposed to MPEG-7 [31]. For envisaged application of the future standard MPEG-7 [12], the scene illumination can be a useful feature. The first MPEG-7 application regarded here is the storage and retrieval of video sequences in data bases. Here, the presence or absence of cast shadows or shading is a useful feature to identify scenes. As shown in Fig. 1, two quite similar video sequences - if only motion and shape is regarded - can be easily identified, if illumination is regarded too. A second MPEG-7 application is semi-automated multimedia editing. Here, the illumination in scenes of different video sources may be different. If the illumination as feature of the scene was known, an illumination harmonization process [29] could be started automatically. A third MPEG-7 application is performing arts, where the scene illumination can be a useful feature to select video scenes.

In this report, the scene illumination is proposed as feature for video sequences to assist MPEG-7 video indexing applications. More specific, the proposed feature will be the presence and strength of three *temporal* illumination effects. The first effect is temporal changing object shading. Temporal changing shading causes a variation of the measured image signal of an object surface due to rotation or translation of the object. The second effect is moving cast shadows. A moving cast shadow causes variations in the measured image signal of an object surface due to another object that moves between the surface and a light source. The third effect is temporal changes in global illumination intensity.

The proposed illumination descriptor will be based on temporal luminance variations along object motion trajectories. Thus, motion information like displacement vector fields (DVF) from an MPEG-4 coded video is necessary additionally to the video sequence. The

(a) (b)

FIG. 1 – *Two typical head-and-shoulder sequences (a) "Akiyo" and (b) "Erik": The two scenes are quite similar regarding only motion or shape of the moving objects. Regarding additionally illumination, the sequences show marking differences: In (a), no illumination effects are present. Whereas in (b), cast shadows and object shading are present.*

descriptor will be a single scalar that can be calculated hierarchically, i.e. for an image region, for a single image or for an image sequence. The descriptor will give a high score if shading, cast shadows or global illumination changes are strong or cover large areas in the considered video.

The report will further develop a method for the extraction of the descriptor from video. It will be shown that the extraction of the proposed descriptor describing *temporal* illumination effects can be fully automatic. This is of high importance for its practical use. To the contrary, descriptors of static illumination effects or descriptors of light sources are restricted either to unicolored scene background [7][24], or to unicolored objects [21][9], or they need additional input data as 3D object shape [13][27][29]. The proposed descriptor extraction method shall consider errors in all used data, i.e. the video images *and* the used motion information.

This report is organized as follows. Section 2 introduces the proposed illumination effect descriptor. Then, Section 3 presents a method for its robust and automatic extraction from video. The performance of the extraction method is investigated in Section 4 and Section 5 presents sample retrieval results. Finally, Section 6 gives a conclusion.

2 An illumination effect descriptor

2.1 Introduction

In the following, a descriptor for illumination effects in video sequences will be developed. The detection and measure of common illumination effects like shading and cast shadows in a *single* image is a hard problem. Classical references addressing shading and cast shadows in a single image restrict the scene either to unicolored objects [21][9] or unicolored background [7][24]. In [24] even a known background image without objects is required.

Such restrictions are not practical for real video material. An illumination effect descriptor is useful only, if a wider range of video material can be addressed. Further, an automatic extraction of the descriptor should be possible. This led to the idea of this report to measure

temporal variations of the image signal due to illumination effects. The evaluation of temporal changes does not require the mentioned restrictions for objects or background. Also, measuring temporal changes of shading or cast shadows on surfaces of moving objects has been shown to be successful for illumination estimation [13][27][29], shape estimation [22], motion estimation [19][30][32] and shadow detection and tracking [28].

2.2 Definition of the descriptor

For the present, the descriptor is defined for a single picture element (pel) at the 2D image position $\mathbf{p}$ in an image s_k, where s indicates the image luminance and k the time instant. Later, the descriptor may be averaged over a set of pels in a single video image or in a sequence of video images. The descriptor for a single pel is defined using three elements of data:

1. the image luminance $s_k(\mathbf{p})$ in the current image,
2. the displacement vector $\mathbf{d}(\mathbf{p})$ pointing from the position $\mathbf{p}$ at time instant k to the corresponding position in the preceding image, and
3. the corresponding image luminance $s_{k-1}(\mathbf{p}+\mathbf{d}(\mathbf{p}))$ in the preceding image.

To describe the temporal change of the image luminance of a pel due to illumination effects, several models have been proposed in the literature. The temporal change is described either by an additive constant [10][20][3] or by a general linear model (a factor and an additive constant) [19][4][18][36]. The parameters of the signal models for neighboring image positions may be spatially constrained [10][20][3][19]. All these models assume matte and opaque object surfaces.

In this report, the linear description by a factor has been chosen. This reflects best the physical influence of illumination changes on the luminance signal. In fact, this factor is the displaced frame ratio (DFR)

$$dfr(\mathbf{p}) \;=\; \frac{s_k(\mathbf{p})}{s_{k-1}(\mathbf{p}+\mathbf{d}(\mathbf{p}))}\,, \tag{1}$$

that describes the luminance variation of a pel along its motion trajectory. To show the physical meaning of the DFR, a luminance signal model is introduced. This model is based on the following assumptions:

- Radial photometric distortions of perspective camera projection [5] can be neglected.
- The gamma nonlinearity of the video camera can be neglected.
- All light sources in the scene have the same spectrum (color) [29].
- The object surfaces in the scene are Lambertian.
- Camera noise can be neglected.

The model describes the image luminance

$$s_k(\mathbf{p}) \;=\; E_k(\mathbf{p})\;\rho_k(\mathbf{p}) \tag{2}$$

by the product of the irradiance $E_k(\mathbf{p})$ and reflectance $\rho_k(\mathbf{p})$ of the object surface. The irradiance is the received light power per receiving object surface. The reflectance is the ratio between reflected and received light power. Inserting the luminance signal model from Eq. 2 into Eq. 1 leads to the frame ratio

$$dfr(\mathbf{p}) \;=\; \frac{E_k(\mathbf{p})}{E_{k-1}(\mathbf{p}+\mathbf{d}(\mathbf{p}))}\;\frac{\rho_k(\mathbf{p})}{\rho_{k-1}(\mathbf{p}+\mathbf{d}(\mathbf{p}))}\;. \tag{3}$$

Assuming perfect motion compensation, the same scene detail is visible at two corresponding image positions $\mathbf{p}$ and $\mathbf{p}+\mathbf{d}(\mathbf{p})$ and thus, $\rho_{k-1}(\mathbf{p}+\mathbf{d}(\mathbf{p})) = \rho_k(\mathbf{p})$ holds. Then, the DFR simplifies to

$$dfr(\mathbf{p}) \;=\; \frac{E_k(\mathbf{p})}{E_{k-1}(\mathbf{p}+\mathbf{d}(\mathbf{p}))} \tag{4}$$

and describes the temporally changing irradiance due to shading on moving object surfaces, due to moving cast shadows or due to global illumination intensity changes. These illumination effects can cause the DFR to be larger or smaller than one. Therefore, the single-pel-descriptor is defined to be the modified DFR according to

$$\tilde{dfr}(\mathbf{p}) \;=\; \begin{cases} dfr(\mathbf{p}) & \text{if } dfr(\mathbf{p}) \;\geq\; 1 \\ 1/dfr(\mathbf{p}) & \text{if } dfr(\mathbf{p}) \;<\; 1 \end{cases}\;. \tag{5}$$

By the way, Eq. 5 ensures also that the illumination descriptor is independent from the sense of the video sequence (forwards or backwards).

The descriptor proposed in this report is the mean of the single-pel-descriptor for a certain set $\mathcal{Z}$ of pels, i.e. an image region, an image or an image sequence. The descriptor is therefore defined as

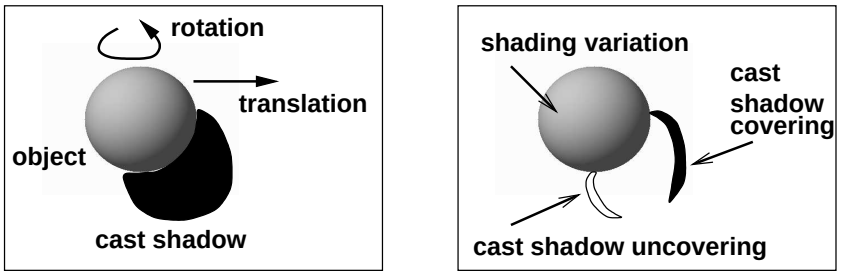
$$D \;=\; \frac{1}{|\mathcal{Z}|}\sum_{\mathbf{p}\in\mathcal{Z}} \tilde{dfr}(\mathbf{p})\;. \tag{6}$$

E.g. a descriptor value of $D = 1.02$ means that the illumination effects cause image luminance changes from one image to the next by in the mean 2% of the image luminance.

2.3 Discussion of the descriptor

In Fig. 2, a typical scene with a shaded moving object and a moving cast shadow in the background is shown. The proposed illumination effect descriptor is based on luminance variations along motion trajectories. The luminance variation and thus the descriptor will be influenced by properties of the scene illumination like:

- **Light source intensity:** A stronger light source will cause a stronger contrast in object shading. If the object moves, there will be stronger luminance variations along



(a) (b)

FIG. 2 – *(a) Typical scene of a shaded object with a cast shadow in the background and (b) image signal changes caused by object translation and rotation*

the motion trajectories. Further, a stronger light source will cause shadows with stronger contrast that causes stronger luminance variations in regions of uncovering and covering cast shadows.

– **Light source position:** Depending on the position of light sources, different parts of object surfaces will be influenced by shading. Further, the position, size, shape and contrast of cast shadows is influenced.

Unfortunately, the descriptor will be influenced also by properties of the scene geometry like:

– **Object motion:** The amount of rotation of an object influences the temporal changes of the shading on its surface, whereas the amount of its translation influences the size and shape of image regions that are covered or uncovered by the cast shadow(s) related to this object.
– **Scene geometry:** The distance between objects in the scene and their shape have influence on the position, size, shape and intensity of cast shadows.

To sum up, the descriptor reflects all those physical scene parameters that affect the temporal variation of illumination effects. This temporal variation depends mainly on light source intensity and object motion. Static illumination effects like not moving shaded objects or like not moving cast shadows are not covered by the descriptor developed in this report.

3 Method for automatic descriptor extraction from video

This section describes an automatic, noise-adaptive, robust method to derive the descriptor automatically from a video sequence. The single scalar descriptor can be evaluated for an arbitrarily image region, for one whole image or for a sequence of images. In this

section, it is assumed that the descriptor is evaluated for an arbitrarily image region in one image of an image sequence. The calculation of the descriptor needs as input the current image, the preceding image and a vector field for the displacements (DVF) between both images.

In Section 3.1, the method for robust descriptor extraction is presented. It exploits only pels with a smooth displaced frame ratio (DFR) to be robust against motion compensation errors. The smoothness constraint is noise adaptive. It is based on a camera noise variance estimate that described in Section 3.2. The noise variance estimator is extracted from the displaced frame difference (DFD) and is itself robust against motion compensation errors. It exploits only pels with small motion compensation errors. Therefore, Section 3.3 presents an estimator of the variance of the DFD as it would be in case of no motion compensation errors.

3.1 Calculation of illumination effect descriptor

In the following a method for the calculation of the descriptor for a set $\mathcal{Z}$ of pels will be introduced. Assuming smooth shaped object surfaces, the idea is that the displaced frame ratio (DFR) from Eq. 4 should be smooth if the DFD is caused by shading, cast shadows or global illumination intensity changes. Regarding camera noise, the DFR may vary "a little" but not "too much". An image region having a spatially non-homogeneous DFR is assumed either to contain other illumination effects than the considered ones or, notably, it is assumed to be badly motion compensated.

To consider these cases, the descriptor is extracted only from a subset $\mathcal{B}$ of pels instead of from all $\mathcal{Z}$ as defined for the descriptor in Eq. 6. This subset is defined by

$$\mathcal{B} = \{\mathbf{p} \in \mathcal{Z} \mid \sigma_{DFR,local}(\mathbf{p}) < th_{DFR} \wedge s_k(\mathbf{p}) \geq s_{min}\} , \tag{7}$$

where $\sigma_{DFR,local}(\mathbf{p})$ is the variance of the DFR calculated in a local neighborhood of $\mathbf{p}$ and s_{min} a threshold discussed further down. By thresholding the local variance of the DFR, $\mathcal{B}$ retains pels of smooth DFR. The DFR will be smooth in image regions with

- precise motion compensation,
- smooth shape of the objects in the scene, and
- Lambertian object surfaces.

Perfect motion compensation ensures the validity of Eq. 4 and smooth object shape and a Lambertian object surface causes a smooth irradiance in Eq. 4. The threshold th_{DFR} is chosen such, that it equals the variance of the DFR due to camera noise, only. Adding camera noise to the luminances in Eq. 1, then assuming the noise to be small compared to the luminances, further calculating the variance of the DFR in Eq. 1 and finally replacing the luminance $s_k(\mathbf{p})$ by the estimate

$$\hat{\mu}_s = \frac{1}{|\mathcal{Z}|} \sum_{\mathbf{p} \in \mathcal{Z}} s_k(\mathbf{p}) \tag{8}$$

of the mean of the image luminance leads to the threshold

$$th_{DFR} \;=\; \frac{\hat{\sigma}_s^2}{\hat{\mu}_s^2} \tag{9}$$

with $\hat{\sigma}_s$ a camera noise variance estimate that will be described in the following section.

The second constraint in Eq. 7 applying a threshold s_{min} is heuristic and tries to exclude dark image regions where the calculation of the DFR is unsure due to quantization and noise. It is

$$s_{min} \;=\; min\left\{\frac{2}{D'-1}\,,\,128\right\} \tag{10}$$

with D' being a guess of the descriptor D. Note that descriptor values for typical images are close to and superior to one. In this report, D' is derived by repeating the estimation: In a first estimation, s_{min} is set to zero and in the second estimation, the result of the first estimation is used as D'.

3.2 Estimation of the variance of camera noise

To estimate the variance of camera noise, the displaced frame difference (DFD) is described by

$$dfd(\mathbf{p}) \;=\; s_k(\mathbf{p}) + \epsilon_k \;-\; s_{k-1}(\mathbf{p}+\mathbf{d}(\mathbf{p})) - \epsilon_{k-1}\;, \tag{11}$$

where ϵ_{k-1} and ϵ_k account for camera noise in the two succeeding images s_{k-1} and s_k, $\mathbf{p}$ is a 2D image position and $\mathbf{d}(\mathbf{p})$ is a displacement vector pointing backwards from k to $k-1$. Further, perfect motion compensation is assumed, an assumption that will be justified by a choice of robust pels some paragraphs further down. Then,

$$s_k(\mathbf{p}) \;-\; s_{k-1}(\mathbf{p}+\mathbf{d}(\mathbf{p})) \;=\; 0 \tag{12}$$

holds and

$$dfd(\mathbf{p}) \;=\; s_k(\mathbf{p}) + \epsilon_k \;-\; s_k(\mathbf{p}) - \epsilon_{k-1} \;=\; \epsilon_k - \epsilon_{k-1}\;. \tag{13}$$

Assuming a zero mean stationary camera noise, the DFD variance is

$$\sigma^2_{DFD,robust} \;=\; E\left[dfd^2(\mathbf{p})\right] \;=\; 2\,\sigma_s^2 \tag{14}$$

and thus, the camera noise variance is

$$\sigma_s^2 \;=\; \frac{1}{2}\,\sigma^2_{DFD,robust}\;. \tag{15}$$

Eq. 15 is valid only, if the motion compensation is perfect. Let us assume, that for the pels $\mathbf{p} \in \mathcal{A} \subseteq \mathcal{Z}$ this assumption is valid. Then, Eq. 15 and the assumption of an ergodic DFD leads to the following estimator:

$$\hat{\sigma}_s^2 \;=\; \frac{1}{2\,|\mathcal{A}|} \sum_{\mathbf{p}\in\mathcal{A}} dfd^2(\mathbf{p})\;. \tag{16}$$

The region $\mathcal{A}$ will be defined such that large motion compensation errors are excluded. Large motion compensation errors are assumed to cause a large DFD. By the way, only in this case, motion compensation errors disturb the estimator in Eq. 16. Regions of large motion compensation errors are detected by the following thresholding operation

$$\mathcal{A} \;=\; \{\mathbf{p}\in\mathcal{Z} \;|\; |dfd(\mathbf{p})| \;<\; th_{DFD}\} \tag{17}$$

with th_{DFD} a threshold adaptive to the image and to the resolution of displacement estimation according to

$$th_{DFD} \;=\; \sqrt{\hat{\sigma}^2_{DFD}}\;, \tag{18}$$

where $\hat{\sigma}^2_{DFD}$ is the estimated variance of the DFD for the case where the DFD is caused only by the limited resolution of the motion estimator. A DFD caused by large motion estimation errors is assumed to be stronger. The variance estimator $\hat{\sigma}^2_{DFD}$ will be developed in the following section. By the use of robust pels in $\mathcal{A}$, the noise variance estimator itself is robust against motion compensation errors.

3.3 Estimation of the maximum variance of the displaced frame difference

This section develops an estimator for the variance of the displaced frame difference (DFD) for the imaginary case where the DFD is caused only by the limited resolution of an applied motion estimator. The resolution is denoted as ϵ_{DVF} [pel], e.g. $\epsilon_{DVF} = 1/2$ for half pel resolution. The limited displacement resolution will cause the following DFD

$$dfd(\mathbf{p}) \;=\; s_k(\mathbf{p}) \;-\; s_{k-1}(\mathbf{p}+\mathbf{d}(\mathbf{p})+\epsilon_{DVF}\,\mathbf{e}) \tag{19}$$

in the worst case. Here, s_{k-1}, s_k are two consecutive images, $\mathbf{p}$ is a 2D image position, $\mathbf{d}(\mathbf{p})$ is a displacement vector and $\mathbf{e}$ is a 2D unit vector of arbitrary direction. Image noise is neglected here assuming that noise is smaller than errors caused by the limited displacement resolution. Apart of ϵ_{DVF}, perfect motion compensation according to Eq. 12 is assumed. Inserting Eq. 12 into Eq. 19 leads to

$$dfd(\mathbf{p}) \;=\; s_k(\mathbf{p}) \;-\; s_k(\mathbf{p}+\epsilon_{DVF}\,\mathbf{e})\;. \tag{20}$$

Assuming that the current image s_k is linear in a local neighborhood of radius ϵ_{DVF} around $\mathbf{p}$, the DFD can be described by

$$dfd(\mathbf{p}) \;=\; \epsilon_{DVF} \;\cdot\; (\,s_k(\mathbf{p}) \;-\; s_k(\mathbf{p}+\mathbf{e})\,)\;. \tag{21}$$

The DFD will be maximum, if $\mathbf{e}$ is in the direction of the spatial image signal gradient $\mathbf{g}_s(\mathbf{p})$. Thus, the DFD is limited by

$$|dfd(\mathbf{p})| \leq \epsilon_{DVF} \cdot |\mathbf{g}_s(\mathbf{p})| \tag{22}$$

and the DFD variance will limited by

$$\sigma^2_{DFD} \leq \epsilon^2_{DVF} E\left\{\mathbf{g}_s^T(\mathbf{p})\,\mathbf{g}_s(\mathbf{p})\right\} . \tag{23}$$

A reasonable estimator for the DFD variance caused by the limited resolution of the displacement estimator is then

$$\hat{\sigma}^2_{DFD} = \frac{1}{|\mathcal{Z}|} \sum_{\mathbf{p}\in\mathcal{Z}} \hat{\mathbf{g}}_s^T(\mathbf{p})\,\hat{\mathbf{g}}_s(\mathbf{p}) \tag{24}$$

assuming $|\mathcal{Z}| >> 1$ and with $\hat{\mathbf{g}}_s(\mathbf{p})$ the measured spatial image gradient at image position $\mathbf{p}$.

4 Performance analysis

To analyze the errors of descriptor extraction from video, three experiments have been carried out:

1. Measurement of systematic estimation errors using a synthesized image pair with simulated moving cast shadows on the background
2. Measurement of estimation error variances using a synthesized image pair with simulated cast shadows on the background and with synthesized additive noise
3. Measurement of estimation error variances using a real images pairs with synthesized additive noise, natural motion, natural cast shadows and natural object shading

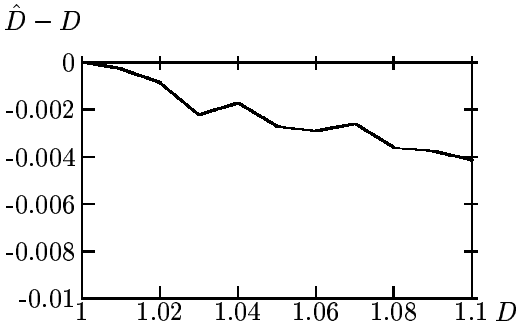
For all experiments, the descriptor has been extracted automatically by the method described in Section 3. The displacement vector field (DVF) needed by the extraction method has been derived using hierarchical block matching with half pel accuracy and cross-correlation-coefficient-criterion [1].

The first experiment uses a synthesized image pair. The first image is the first image of the test sequence "Tai", see Fig. 3(a). The second image is generated from the first one by pel by pel application of a factor either D or $1/D$, where D is the ground truth descriptor value describing illumination changes. This process simulates covering or uncovering by moving cast shadows. Fig. 3(b) shoes the second synthesized image for $D = 1.1$. In image regions marked white in Fig. 3(c), the factor D is applied, in the black regions $1/D$ is applied. The whole image is affected by simulated illumination effects of same strength. The question is, if the algorithm is capable to estimate the factor D.

For the first experiment, the results are shown in Fig. 4. There is always a systematic, negative estimation error that increases linearly with the ground truth descriptor value D,

(a) (b) (c)

FIG. 3 – *Generation of a synthesized image pair (a) and (b) with simulated illumination effects by pel-by-pel multiplication of (c) a simulation pattern: White regions indicate a multiplicative change of image luminance by D (e.g. uncovering by shadow) and black regions a change by $1/D$ (e.g. covering by shadow).*



(a)

FIG. 4 – *Performance analysis: Systematic estimation error $\hat{D} - D$ versus ground truth descriptor value D using synthesized images*

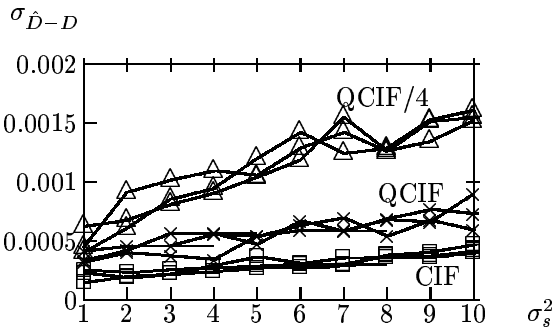


FIG. 5 – *Performance analysis: Standard deviation $\sigma_{\hat{D}-D}$ of stochastic estimation error versus variance σ_s^2 of added camera noise for CIF, QCIF and QCIF/4 image formats and for different ground truth descriptor values $D = 1.04$, $D = 1.07$ and $D = 1.1$ using synthesized images*

but only by a gradient of 1/20. In other words, the estimated descriptor value is always slightly smaller than the true value. The reason for this is the motion compensation by a DVF before calculation of the displaced frame ratio (DFR). The displacement estimation compensates partially also some of the temporal illumination changes, thus the derived descriptor value is always slightly too small. To minimize this effect, it is important to rubustify the displacement estimation against illumination effects. Therefore, different matching criteria as maximum-absolute-difference, mean-squared-error, and cross-correlation-coefficient had been employed in pre-experiments. As can be expected, the cross correlation coefficient gave clearly the best results. All shown experiments have thus been carried out with the cross correlation coefficient criterion. To sum up, the systematic error is in the third digit of the descriptor for typical illumination effects that have a $D \leq 1.1$. Whereas, the relevant digits are the first and second ones (see results in Section 5).

For the second experiment, the results are shown in Fig. 5. The question here is - beside the systematic error measured by the first experiment - how sensible are the results in presence of camera noise. Each curve shows the standard deviation of estimation errors of the descriptor for different variances of added Gaussian camera noise. To calculate a single point of one of the curves, the same experiment has been repeated 25 times with different noise realizations of same variance. The nine curves are different regarding the chosen image format - CIF (352×288 pels), QCIF (176×144 pels) or QCIF/4 (88×72 pels) - and regarding the ground truth descriptor value, $D = 1.04$, $D = 1.07$ or $D = 1.1$. In this experiment, estimation errors are caused by displacement estimation errors due to the added camera noise as well as directly by noise during the calculation of the DFR. It can be seen that the estimation error standard deviation increases with the camera noise. With increasing image format, i.e. increasing number of exploited pels, the error standard deviation goes down.

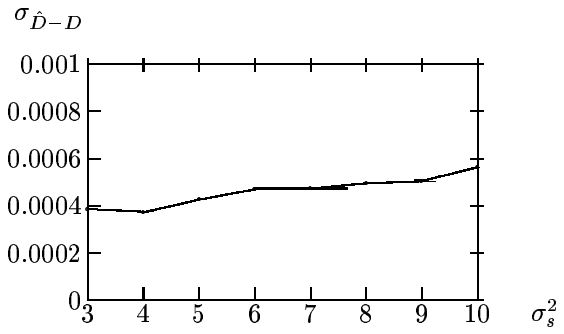


FIG. 6 – *Performance analysis: Results for a real image sequence in CIF format with added synthetic noise: Standard deviation of stochastic estimation error $\sigma_{\hat{D}-D}$ versus variance of added camera noise σ_s^2.*

This is a well known property of estimators that are disturbed by uncorrelated noise. The ground truth descriptor value, i.e. the strength of illumination effects in the video sequence, has nearly no influence for the investigated values of $1 \leq D \leq 1.1$. The stochastic errors (Fig. 5) are generally smaller than the systematic error (Fig. 4), even at a camera noise variance of 10 and with only 6500 exploited pels (QCIF/4).

For the third experiment, the results are shown in Fig. 6. For these experiments, 49 real image pairs from the 50 images of the test sequence "Tai" (first image see Fig. 3(a)) have been used. For each image pair, the descriptor has been extracted by the method from Section 3. This descriptor value is supposed to be the true one. Then, for each image pair the extraction is repeated 25 times by adding to the images different realizations of camera noise of the same variance. The standard deviation refers to the difference between this noisy result and the descriptor value extracted from the original image pair. As can be seen, the error in the descriptor extraction from real images introduced by camera noise has the same variance as in the second experiment for synthetic images (compare Fig. 5 for CIF). In this experiment, the errors are caused by displacement estimation errors due to real *and* added synthetic camera noise as well as directly by noise during the calculation of the DFR.

To sum up, systematic and stochastic errors are inferior to 0.01 in amplitude, i.e. in the 3rd digit of the descriptor. Whereas, for the experiments shown in the next section, the first and second digits are relevant.

5 Results

To obtain the experimental results of this section, the set of ten video image sequences shown in Fig. 7 is used. The set contains head-and-shoulder scenes of moving and talking

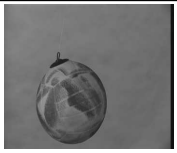
Test Sequence	Format	Test Sequence	Format
 Akiyo	CIF 300 images	 Tai	CIF 50 images
 Claire	CIF 155 images	 Erik	CIF 50 images
 Mother	CIF 300 images	 Table Tennis	SIF Second Part 169 images
 Jürgen	CIF 24 images	 Ball	CIF 80 images
 Christoph	CIF 50 images	 Roma	CIF 24 images

FIG. 7 – *The set of test sequences used as data base for the retrieval experiments*

FIG. 8 – *Results of two image sequence retrievals using the illumination descriptor as only criterion with (a) "Akiyo" and (b) "Table Tennis" as sample image sequences*

persons with and without visible cast shadows or shading effects. Further, a moving toy-car, a rapidly moving toy-ball and a sport scene are included.

For the image sequences, corresponding sequences of displacement vector fields (DVF) are available. The DVFs had been derived by hierarchical block matching with half pel accuracy and cross-correlation-coefficient-criterion [1]. For each image sequence, the descriptor is evaluated for each image, beginning with the second image (since a preceding image is necessary). For each image sequence, the extracted descriptor values are averaged to get a single descriptor value for each image sequence. This single descriptor value is used as index for each image sequence.

In this data base of ten image sequences, search experiments are carried out. It is looked for image sequences that have illumination effects of comparable strength as a sample image sequence. As sample image sequences, "Akiyo" and "Table Tennis" are used. "Akiyo" is a blue-screen sequence news sequence with no visible illumination effects. The extracted illumination effect descriptor value is 1.0010 (a value close to one means few illumination effects). In "Table Tennis", a rapidly moving player casts two shadows that move quickly with him. The extracted illumination effect descriptor value is 1.0237, i.e. in the mean the luminances are changed by 2.37% of their amplitude by illumination effects. The search criterion is the absolute distance between descriptor values using a threshold of 0.01, i.e. 1% of image luminance.

The retrieval for "Akiyo" results in sequences with weak illumination effects, see Fig. 8(a). In "Christoph" and "Jürgen", the persons cast weak cast shadows on the background and their face is slightly shaded. In "Claire", the face is shaded. In "Mother", the hand and the head of the mother casts shadows, but they don't move a lot.

The retrieval for "Table Tennis" results in sequences with strong illumination effects, see Fig. 8(b). In "Tai", the face of the person is visibly shaded by one dominating light source. In "Erik", the persons face is shaded and the person casts a strong cast shadow that moves with the person.

The sequence "Roma" has average illumination effects and is not detected by any of the two retrievals (descriptor value: 1.0128). The sequence "Ball" has very strong illumination effects (shading changes heavily due to about 10 degrees of object rotation from image to image) and is even not detected by the second retrieval (descriptor value: 1.0332).

6 Conclusions

In this report, a descriptor for illumination effects in a video image sequence is proposed. The descriptor is defined to be the average of a single-pel-descriptor for an image region, for a single image or for a sequence of images.

The descriptor is based on the displaced frame ratio (DFR). The DFR is the pel-by-pel ratio of image luminances after motion compensation. A DFR unequal one indicates a multiplicative change of image luminance due to temporally changing shading on object surfaces caused either by object rotation under directed light, by moving cast shadows or by global illumination changes. The single-pel-descriptor is defined to be the DFR, where DFR values smaller than one are replaced by their inverse values. The descriptor describes *temporal changing* illumination effects and is strongly influenced by light source intensities, by light source positions, by motion of the objects and by the scene geometry.

Further, this report presents a method for the noise-robust and automatic extraction of the descriptor from video. Therefore, only pels in image regions of spatially smooth DFR are considered. Hereby, regions of large motion compensation error or large camera noise are excluded. A performance analysis of the extraction method reveals that extraction errors are more than 10 times smaller than relevant descriptor value changes from scene to scene.

The descriptor has been used for indexing of a set of 10 image sequences of simple complexity (indoor scenes of persons, head-and-shoulder scenes). The retrieval results show that the illumination effect descriptor is precise and sensible enough to distinguish between video sequences with different scene illumination. The experiments prove that the scene illumination as new physical scene property is useful for video indexing.

Nevertheless, the proposed descriptor has an inherent inconvenience. Whereas with "Akiyo" the absence of illumination effects is very well reflected by the descriptor value, the quantitative description of illumination effects in some other scenes seems to be over- or underestimated. For example in "Christoph" and "Jürgen", the illumination effects seem to be somehow underestimated. This is because they do not change a lot temporally and result thus in a weaker descriptor value. On the other hand, the descriptor gives a strong value for the sequence "Ball" which seems to be an overestimation. The ball is only weakly shaded but it rotates so quickly that this causes a lot of signal changes due to shading. The inconvenience is the inherent dependency of the presented illumination descriptor on the motion of objects. It may be avoided in future by a combination of illumination and object motion descriptors.

Références

[1] M. Bierling, Displacement estimation by hierarchical blockmatching, 3rd SPIE *Symposium on Visual Communications and Image Processing*, Cambridge, USA, November 1988, pp. 942-951.

[2] R. Brunelli, O. Mich, C.M. Modena, A survey on the automatic indexing of video data, *Journal of Visual Communication and Image Representation*, Vol. 10, No. 1, March 1999, pp. 78-112.

[3] M. Gilge, Motion estimation by scene adaptive block matching and illumination correction, *Image Processing Algorithms and Techniques*, SPIE Vol. 1224, 1990, pp. 355-366.

[4] F.J. Hampson, R.E.H. Franich, J.C. Pesquet, J. Biemond, Pel-recursive motion estimation in the presence of illumination variations, *ICIP'96 International Conference of Image Processing*, Lausanne, Switzerland, 16.-19.9.1996, Vol. 1, pp. 101-104.

[5] B.K.P. Horn, *Robot vision*, MIT Press, Cambridge, 1987.

[6] M. Irani, B. Rousso, A. Peleg, Computing Occluding and Transparent Motions, *International Journal of Computer Vision*, Vol. 12, No. 1, January 1994, pp. 5-16.

[7] C. Jiang, M.O. Ward, Shadow segmentation and classification in a constrained environment, *CVGIP: Image Understanding*, Vol. 59, No. 2, March 1994, pp. 213-225.

[8] G.J. Klinker, S. A. Shafer, T. Kanade, A Physical approach to color image understanding, *International Journal of Computer Vision*, Vol. 4, No. 7, July 1990, pp. 7-38.

[9] C.H. Lee, A. Rosenfeld, Improved methods of estimating shape from shading using the light coordinate system, in B.K.P. Horn, M.J. Brooks (Ed.), *Shape from shading*, MIT Press, Cambridge 1989, pp. 323-347.

[10] C. R. Moloney and E. Dubois, Estimation of motion fields from image sequences with illumination variation, *ICASSP*, Toronto, Canada, 14-17 May 1991, Vol. 4, pp. 2425-2428.

[11] MPEG Requirements Group, *Context and Objectives*, MPEG Document ISO/IEC JTC1/SC29 WG11 N2460.

[12] MPEG Requirements Group, *MPEG-7 Applications*, MPEG Document ISO/IEC JTC1/SC29 WG11 N2462.

[13] N. Mukawa, Estimation of light source information from image sequence, *Systems and Computers in Japan*, Vol. 23, No. 10, October 1992, pp. 92-99.

[14] H. Murase, Surface shape reconstruction of un undulating transparent object, *IEEE International Conference on Computer Vision and Pattern Recognition*, 1991, pp. 313-317.

[15] H. Murase, S.K. Nayar, Illumination planning for object recognition using parametric eigenspaces, *IEEE Trans. on Pattern Recognition and Machine Intelligence*, Vol. 16, No. 12, December 1994, pp. 1219-27.

[16] S. K. Nayar, R. M. Bolle, Reflectance based object recognition, *International Journal of Computer Vision*, Vol. 17, No. 3, March 1996, pp. 219-239.

[17] S.K. Nayar, K. Ikeuchi, T. Kanade, Surface reflection: Physical and geometrical perspectives. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, Vol. 13, No. 7, July 1991, pp. 611-634.

[18] S. Negahdaripour, C.-H. Yu, A generalized brightness change model for computing optical flow, *International Conference on Computer Vision*, 11.-14. May 1993, Berlin, pp. 2-11.

[19] H. Nicolas, C. Labit, Motion and illumination variation estimation using a hierarchy of models: Application to image sequence coding, *Journal of Visual Communication and Image Representation*, Vol. 6, No. 4, December 1995, pp. 303-316.

[20] A. Nomura, H. Miike, K. Koga, Determining motion fields under non-uniform illumination, *Pattern Recognition Letters*, Vol. 16, No. 3, March 1995, pp. 286-296.

[21] A.P. Pentland, Finding the illumination direction, *Journal of the Optical Society of America*, Vol.. 72, No. 4, April 1982, pp. 448-455.

[22] A.P. Pentland, Photometric Motion, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 9, September 1991, pp. 879-890.

[23] Y. Rui, T.S. Huang, S.-F. Chang, Image retrieval: Current techniques, promising directions, and open issues, *Journal of Visual Communication and Image Representation*, Vol. 10, No. 2, June 1999, pp. 39-62.

[24] G.G. Sexton, X. Zhang, Supression of shadows for improved object discrimination, *IEEE Colloquium on Image Processing for Transport Applications*, London, UK, December 1993, pp. 9/1-9/6.

[25] S. A. Shafer, *Shadows and silhouettes in computer vision*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1985.

[26] T. Sikora, MPEG-7 visual descrptors - an overview, W*IAMIS '99*, European Workshop on Image Analysis for Multimedia Interactive Services, 31.5.-1.6.1999, Berlin, Germany.

[27] J. Stauder, Estimation of point light source parameters for object-based coding, *Signal Processing: Image Communication*, Vol. 7, No. 4-6, November 1995, pp. 355-379.

[28] J. Stauder, R. Mech, J. Ostermann, Detection of moving cast shadows for object segmentation, *IEEE Trans. on Multi Media*, Vol. 1, No. 1, March 1999, pp. 65-76.

[29] J. Stauder, Augmented Reality with Automatic Illumination Control Incorporating Ellipsoidal Models, *IEEE Trans. on Multi Media*, Vol. 1, No. 2, June 1999, pp. 136-143.

[30] J. Stauder, Object rotation axis from shading, *ICASSP'99*, Phoenix, Arizona, USA, 15.-19.3.1999, Vol. 6, pp. 3285-3288.

[31] J. Stauder, *An illumination effect descriptor for video image sequences*, MPEG-7 proposal No. P099, Simulation Group Adhoc Meeting, Lancaster, Great-Britain, February 1999.

[32] J. Stauder, H. Nicoloas, Motion-based video indexing evaluating object shading, *ICIP'99 International Conference on Image Processing*, Kobe, Japan, 24.-28.10.1999.

[33] J. Stauder: "An Illumination Effect Descriptor for Video Sequences", submitted to the Journal of Visual Communication and Image Representation, October 1999.

[34] J. Stauder: "A video descriptor based on illumination effects", submitted to ECCV-2000.

[35] P. Treves, J. Konrad, Motion estimation and compensation under varying illumination, *IEEE International Conference on Image Processinaag IPIC*, Austin, Texas, USA, 13.-16. November 1994, pp. 373-377.

[36] J. Wei, Z.-N. Li: Motion compensation in color video with illumination variations, *ICIP'97 International Conference on Image Processing*, 26-29.10.1997, Santa Barbara, USA, pp. 614-617.

[37] S. Yi, R.M. Haralick, L.G. Shapiro, Optimal sensor and light source positioning for machine vision, *Computer Vision and Image Understanding*, Vol. 61, No. 1, January 1995, pp. 122-137.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399